# Reinforcement Learning

## Scenario Based MCQs for End Course Test

**Question1:** Smart-Pets is a startup that facilitates the owners of pets like cats and dogs to interact with the animals online before adopting them. They use a sophisticated IT infrastructure that enhances the interaction. As of now they have 15 cats and 7 dogs available for adoption. They have a user base of 125 customers. However, the number of combinations are huge. They would like to design a solution that will minimize the number of interactions for users so that the entire process doesn’t become cumbersome. Which algorithm would be the best for such a situation?

1. Deep Q Learning
2. Temporal Difference
3. Monte Carlo
4. Dynamic Programming

**Answer:** Option 4. Since we are only interested in identifying the total number of animals that one user should interact with and not the identification of the correct animal for adoption, Dynamic Programming will help minimize the number of interactions.

**Question 2:** Breach Candy hospital has launched a program to help the new doctors (< 5 years of practicing experience) provide more accurate healthcare services. For this they have made an application that suggests the appropriate treatment plans for their patients based on historical data. They believe Monte Carlo methods are the right solution for this situation. Do you agree? What are your reasons?

1. Yes. Monte Carlo methods simulate a variety of combinations and hence cover a lot of variance which will give most accurate results.
2. No. Contextual Bandits will be the best solution as treatment plans differ from patient to patient.
3. Partially Agree. Monte Carlo must be combined with Value Iteration to find the best possible solution.
4. No. Dynamic Programming will give the most accurate results.

**Answer:** Option 2.

**Question 3:** John and Tina are a couple and love each other dearly. However, due to their job requirements they have to be geographically separated for 6 months. Both of them stay worried about what the other had for lunch / dinner as John lives in Jakarta and he hates sea food whereas Tina lives in Hong Kong and hates Chinese food. They talk on the phone daily but never tell the truth about what they had for lunch / dinner so as to avoid each other getting worried. But based on their mood the other can make an estimate whether they liked what they ate or not. Below is a snapshot of their situation for a simpler understanding.

|  |  |  |
| --- | --- | --- |
| **Name / Food** | **Sea Food** | **Chinese** |
| **John** | Sad | Happy |
| **Tina** | Happy | Sad |

You are supposed to suggest which method is best suited to help John and Tina estimate with maximum precision what food the other had on any given day.

1. Upper Confidence Bound Selection
2. Policy Gradients
3. Bellman Equation
4. Markov Chains

**Answer 3:** Option 4. Since food is not the only factor that can cause John and Tina to be happy or sad, this calls for markov chains that include transition probabilities as well as hidden states.

**Question 4:** MG Hector wants all of its cars to learn to continuously drive at the speed of 30 km / hour irrespective of road conditions, flat tire or traffic signal without creating any mortal danger. This is to ensure the car’s ability to save the passengers’ lives in critical situations. Which algorithm is best suited for such scenario?

1. Deep Q Learning Networks because a self-driving car has to take millions of decisions
2. Deep Q Learning networks with Monte Carlo simulations to estimate the values of future actions also
3. Policy gradients as the best state – action pair has to be identified
4. Actor – Critic because the balance between value maximization and policy control is required

**Answer:** Option 3. If you look closely, in this situation the car has just 1 objective – to drive continuously without hitting anything / anyone. Rest all states and actions need not be optimized. So simple usage of policy gradients will do the job.

**Question 5:** You are required to help neurologists at the biology and neuroscience division of Neuralink that is on the verge of creating a cure for Alzheimer disease by helping them choose a combination of RL algorithms to ensure success rate. Which one do you propose?

1. Actor – Critic, Deep Q Networks and Contextual Bandits
2. Contextual Bandits and Deep Q Networks with Replay Buffers
3. Contextual Bandits, Policy Gradients and Deep Q Networks with Replay Buffers
4. Contextual Bandits and Markov Chains

**Answer:** Option 3. Contextual Bandits will give most important contexts. Policy Gradients will choose the state – action pairs with highest values and Deep Q Networks with Replay Buffers will memorize the past actions and their rewards. All these three algorithms contribute directly to information storage in relation to context or values or rewards. Hence, this combination has the highest success rate.

**Question 6:** The city of New York is planning to implement Reinforcement Learning in the electricity generation power house that has 150 turbines that fuel electricity to 769 grids in the city. Their objective is to let the system learn and divide the power generated by the turbines based on consumption, load on each turbine, resources (raw material required to generate electricity) availability, maintenance schedules and predicted power consumption. Which algorithm is best suited for this situation?

1. Dyna - Q
2. SARSA
3. Monte Carlo
4. DQN

**Answer:** Option 4. Deep Q Learning is best suited because exploration is required to cover the distribution for predicted power consumption and there are millions of combinations possible.

**Question 7:** Consider the previous situation. What will be the reward function in the DQN algorithm used?

1. Reduction in load as compared to previous time step
2. Optimal consumption of resources at each time step
3. Exactly matching power generated and power required at each time step
4. Matching power generation with power required at every subsequent time step even if current power generation is lower than required in the current time step

**Answer:** Option 1. If the load in every subsequent time step is reduced then other parameters are automatically optimized.

**Question 8:** Open AI is at the forefront of Reinforcement Learning. The retail giant – Walmart has approached Open AI to help them use reinforcement learning so that robots can train themselves in order to do various tasks such as warehouse management, protecting theft, performing sales, inventory management and optimizing parking facility at their store in Wyoming. Which algorithm / combination of algorithm will you use to achieve this objective?

1. Dynamic Programming
2. Actor – Critic and Deep Q Learning Networks
3. Deep Q Learning with Replay Buffers
4. Policy Evaluation and Expected Value Maximization using Bellman Equation

**Answer:** Option 2. The robot will record the video of each action it performs and use that back as training data for DQN to train using Actor – Critic algorithm for balancing the exploration and value maximization.

**Question 9:** NLP is a domain already dominated by Deep Learning algorithms such as RNN and LSTM. Duck Duck Go needs your help in building an AI Bot that continuously learns from the user’s interaction with the internet in terms of searches made and generates a summary of the relevant web searches as a part of their plan to enhance customer delight. How would you accomplish this task?

1. Use contextual bandits to identify the context behind the search keyword used and then use Deep Q Networks to learn the best words to be used for creating the summary.
2. Use Q Learning to identify the best keywords with highest index values and use LSTM to generate appropriate text.
3. Use policy gradients to learn the process by which a user makes a search (optimal state – action pair) and then use Dynamic Programming to generate summarized text.
4. Use Deep Q Networks to identify the best keywords to be used then train the agent to use those keywords in context using contextual bandits and then use Actor – Critic to analyze the choice of words in generating the summary.

**Answer:** Option 4 requires a lot of data to train but gives the most accurate results.

**Question 10:** Amazon is already conducting A / B testing using contextual bandits to maximize the click through rate of the customers with some success. However they have found that the customers repeatedly visit the same page many times before actually going through with the purchase. As a data scientist what does this mean for you?

1. Contextual Bandits are not trained properly, hyperparameter tuning is required
2. Policy Gradients are a better choice in this situation
3. The maximization of CTR has to be introduced in an MDP because Contextual Bandits treat every visit as a unique visit by a customer
4. Keep increasing the training data to eliminate sampling inefficiency and bias. The learning will converge.

**Answer:** Option 3. MDP will account for interaction effect between the customers over longer period as compared to Epsilon Greedy method that ignores the interaction.

**Question 11:** An NGO is using Temporal Difference Learning method to optimize the transportation of students under 1100 schools but there is a lot of uncertainty in the model. They have used 40 Billion records to train the agent with enough hardware support (NVidia GTX Titan 1080i on a Blade Server with 4TB of RAM and used in 8 couplings). The alpha value is 0.5, gamma value is 0.5 and theta value is 0.0001. They need your help in reducing the uncertainty and stabilizing the learning. What will you do?

1. Reduce alpha to 0.2 and increase gamma to 0.9
2. Increase alpha to 0.9, increase gamma to 0.9 and keep theta constant
3. Reduce alpha to 0.05 and increase theta to 0.001
4. Increase alpha to 0.7 and decrease gamma to 0.3

**Answer:** Option 1. The concept of Stochastic Gradient Descent Applies. Learning becomes smooth when alpha is reduced and when gamma is increased future steps become more and more important which will lead to better planning.

**Question 12:** Dubai Police is training highway patrol cars to be automated and intercept high speed racers. The reward is linked to a successful intercept of the speeding car where success means the racer was stopped but did not face any medical / health damage. They are facing a challenge – what if there is a situation where if the car stays the course and hits five pedestrians – all adults and if it swerves it hits one pedestrian – a child. Should the vehicle swerve or stay its course? There is no clear answer here so they want the agent to come up with a solution and then a decision could be made. Which method is an absolute necessity in this situation?

1. Actor - Critic
2. TD Learning
3. Monte Carlo with bellman equation for value expectation
4. Deep Q Learning with Policy Gradients

**Answer:** Option 3. Monte Carlo will allow the agent to simulate the future rewards without actually exploring the state space and bellman equation will maximize the value thereby generating optimal policy.